**Rebuilding a Community: Data Files and R Scripts Directory**

**Created:** 05/12/2020

**Lead Researcher:** Gabriel Etaat

**Thesis Adviser:** Dr. Robert Heilmayr

**Files:**

1. fire\_data.csv
2. county\_permits.csv
3. housing\_units\_csac.csv
4. bea\_medinc.csv
5. thesis\_data.r
6. thesis\_analysis.r
7. fire\_data.csv

**Purpose:** To compile data on the 26 most destructive wildfires in California between 1998-2018 to be use in the regression analysis model of this thesis. Each fire destroyed at least 120 structures, with at least 80 destroyed in a single county-year pairing.

**Source:** 22 distinct web sources with 38 references were cited in the compilation of this data. This included county fire reports, public insurance records, fire metrics from CAL FIRE, and other trusted primary or secondary sources. A full list of URL’s can be found on the initial excel workbook where the data is listed, fire\_data\_x.xlsx

**Variables and Observations:** 25 unique county-year pairings within the framework of the research population serve as the observations of this study. 6 variables were included in the uploaded csv file from this dataset:

1. county: One of 13 counties where fires observed in the study destroyed a significant number of residential structures (as stated in writing or reported metrics) and a quantifiable number of structures destroyed.
2. year: The year in which a given county was affected by fires observed in the study.
3. fire\_name: The name or names of fires/complex fires that take place for a given county-year pairing. Names were repeated for fires the destroyed a significant and quantifiable number of structures in multiple counties.
4. county\_code: The geocode affiliated with a given county.
5. struc\_dest: The number of structures destroyed by observed fires in a given county a year. Determined by primary fire reports and CAL FIRE fire metrics. Secondary sources such as newspapers were cited if deemed reliable and primary documents were not accessible.
6. struc\_dam\_dest: The number of structures damaged and destroyed by observed fires in a given county a year. Determined by primary fire reports and CAL FIRE fire metrics. Secondary sources such as newspapers were cited if deemed reliable and primary documents were not accessible. Data for this variable was left blank if no reliable sources could be found.

**Thesis Analysis Application:** struc\_dest variable was used as an explanatory variable in all regression models. county and year used for pairing and fixed effects in the merging and analysis of data.

1. county\_permits.csv

**Purpose:** To correlate a variety of independent variables and observe how permitting trends change across regions and years as a result of these factors.

**Source:** The State of the Cities Data Systems (SOCDS), a public data source provided by the U.S. Department of Housing and Urban Development. As described on their webpage, the “Building Permits Database” from which data is obtained “contains data on permits for residential construction issued by about 21,000 jurisdictions collected in the Census Bureau's Building Permits Survey.”

**Download Procedure:**

Step 1: Go to <https://socds.huduser.gov/permits/>

Step 2: Under “MAIN CRITERIA 🡪 “Geography:”, select “States and Counties”

Step 3: Under “MAIN CRITERIA 🡪 “Periodicity:”, select “Annual”

Step 4: Under “Select State(s):”, select “California” and check off “State Total”

Step 5: Click on “Show Counties and Jurisdictions for Selected States”

Step 6: Under “Select Counties”, select all counties and check off “County Total”

Step 7: Under “Select Year(s):”, select years 1998-2018

Step 8: Under “Select Series:”, select all options

Step 9: Click on “Get Data”

Step 10: On the resulting webpage, copy/paste the contents of each table in an Excel sheet with columns C-W pertaining to years 1998-2018, column A pertaining to the permit type, and column B pertaining to the county for each grouping of 6 counties

Step 11: Save the resulting file in a csv format

**Variables and Observations:** The data file contains 354 observations across 23 variables. Each observation pertains to one of 6 permit types: “Total Units”, “Unites in Single-Family Structures”, “Units in All Multi-Family Structures”, “Units in 2-unit Multi-Family Structures”, “Units in 3- and 4-unit Multi-Family Structures”, and “Units in 5+ Unit Multi-Family Structures”. These permit type categorizations are repeated for all 58 counties, plus a 59th California total categorization. Beyond the first to columns defining the regions and permit type, years 1998-2018 are listed to define the total permits issued in a given region-year pairing for a specific permit type.

**Thesis Analysis Application:** After reformatting the data structure to a longer format, the permits issued in a given county and year is analyzed as the dependent variable for the regression analysis. Total units permitted is regressed in the primary model, with single-family and multi-family units permitted regressed in variations of this model to understand the structural typing of houses rebuilt as the direct result of wildfire destruction in subsequent years.

1. Housing\_units\_csac.csv

**Purpose:** To depict the existing housing trends in the state of California and compare them to fluctuations in permitting practices from year-to-year.

**Source:** The California State Association of Counties (CSAC), a lobbying and services organization that serves to assist and represent county governments in the California Legislature. They provide a variety of datasets for public use, including the “Housing Units – Number of Single-Family, Multi-Family, and Mobile Homes by county – 1991 to 2018” database. All data for this specific source was originally accessed from the State of California Department of Finance.

**Download Procedure:** The data, as it appears in the csv file, can be downloaded from the following webpage under DataPile 🡪 Housing Units – Number of Single-Family, Multi-Family, and Mobile Homes by county – 1991 to 2018 (.xlsx)”

**Variables and Observations:** The dataset consists of 4872 observations across 4 variable. Each row represents a unique pairing of 1 of the 58 California counties, one of 28 years from 1991-2018, and one of three housing types (“Single Family”, “Multi Family”, or “Mobile Homes”). For a given county-year-housing type paring, represented in the first three columns, the fourth column represents the total number of housing units that fit within that categorization.

**Thesis Analysis Application:** Existing housing data was primarily used for the depiction of housing trends and summarizing characteristics of the California housing market prior to the regression analysis. It was tested in variations of regression models and other analytical models, but was removed in favor of more representative variables for the relationships being explored.

1. bea\_medinc.csv

**Purpose:** To incorporate an indicator for the wealth of various regions impacted by wildfires by analyzing per capita income

**Source:** CAINC1 database from the U.S. Bureau of Economic Analysis, and agency of the U.S. Department of Commerce. Data for both income and population was collected by the Census Bureau, details listed on rows 183-188 of the csv file.

**Download Procedure:**

Step 1: go to <https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1#reqid=70&step=1>

Step 2: Click on the “PERSONAL INCOME AND EMPLOYMENT BY COUNTY AND METROPOLITAN AREA” tab, and select the database “Personal Income, Population, Per Capita Personal Income (CAINC1)”

Step 3: Under the new data retrieval tool, select “County” under the “Major Area” tab, then “Next Step 🡪”

Step 4: Under the “State” tab, select “California”, then “Next Step 🡪”

Step 5: Under the “Area/Statistic” tab, select both “California state total” and “All counties in California” for “Area”, “All statistics in table” for “Statistic”, and “Levels” for “Unit of Measure”. Then select “Next Step 🡪”

Step 6: Under the “Time Period” tab, select years 1998-2018, then “Next Step 🡪”

Step 7: On the “Selected Data” tab, click on the “Download” icon and select “CSV” to download the data in a csv format

**Variables and Observations:** The file contains 177 rows of data across 25 variables, with titles, legens, and other components in rows 1-5 and 183-188. The first four variables—GeoFips, GeoName, LineCode, and Description—represent the groupings of the observations by county and data type for subsequent entries. GeoFips and GeoNames are synonymous in that a given code corresponds to one of the 58 counties in California or California itself. Similarly, LineCode and Description both reflect the same indicators for observations where wither population, income, or per capita income is measures. The following 21 columns indicate the data type of a county in a given year. This format is altered in R Script thesis\_analysis.r

**Thesis Analysis Application:** In the primary regression model, counties during years observed fires occur are grouped into one of two brackets based on whether their per capita income is above or below $40,000. This binary variable acts as an indicator to analyze the impacts of county wealth on permitting rates, and more importantly, to indicate if counties exhibit rebuilding behavior in subsequent years with significant differences depending on these wealth differences.

1. thesis\_data.r

This R Script depicts the operations needed to upload, clean, format, and merge the distinct data files into a larger dataset, master\_df. Comments are provided throughout the script depicting the procedures employed to create and save this finalized dataset.

1. thesis\_analysis.r

This R Script depicts the operations needed to subset the merged dataset master\_df into dataframes utilized for graphs, tables, and regression analysis input. The operations to create these objects, are then listed as they pertain to the topics of chapter 4 and 5, respectively. Comments are provided throughout the script depicting the procedures employed to create, summarize, and save finalized models and visuals.